**白皮书**

# 使用硬件提高 NFV 性能


编制

Roz Roseboro
Heavy Reading 高级分析师
www.heavyreading.com


代表

www.mellanox.com

**2015 年 7 月**

# 简介

网络功能虚拟化（NFV）为用于交付电信网络功能的平台带来了转变。由于紧耦合专用硬件和软件呈静态且难以扩展，因此标准化 IT 服务器运行虚拟网络功能（VNF），以增强弹性和可伸缩性。

最常见的做法是添加 Hypervisor 层来管理组成 VNF 的虚拟机 (VM)。为了满足通信服务提供商的需求，这一新的体系架构需要具有弹性和可伸缩性，同时又不能影响专用平台的可靠性和性能。

性能的影响因素有很多：虚拟化本身、网络叠加 (overlay) 的使用以及传输控制协议和互联网协议（TCP/IP）的联网协议。使用软件加速是提高性能的一种方法，例如数据平面开发套件（DPDK）库。另一种方法是硬件加速，它利用 SR-IOV 来支持应用程序直接访问。下一代网卡能够减少与封装技术相关的一些开销，这些技术如 VXLAN 和 GENEVE。远程直接内存访问（RDMA）是可替代 TCP/IP 传输的一个新生事物，它将零复制的概念引入了联网领域。

本白皮书的结构如下：

- **第 II 部分**讨论了转变为 NFV 在可靠性、弹性、可伸缩性和性能方面的影响，包括对性能尤其重要的解释。
- **第 III 部分**解释了服务器性能的不同维度。它解释了在考虑服务器平台时需要考虑原始接口速度及数据包吞吐量的重要性。
- **第 IV 部分**探讨了将用于帮助提高服务器性能的技术。它讨论了如何克服由于计算虚拟化、叠加 (overlay) 网络和 TCP/IP 而引起的弊端。

# 转变为 NFV 带来了新要求

转变为 NFV 需要从使用紧耦合网络功能软件的专用硬件转移到标准 IT 服务器平台。因为服务器并未针对网络功能进行优化，所以它们面临的可靠性、弹性、可伸缩性和性能要求不同于 IT 环境。

## 可靠性

由于运行 VNF 的 IT 服务器没有设计为运营商级别的高可用性，因此 NFV 的重点是从五个九 (99.999%) 的硬件可用性转变为五个九的服务可用性。NFV 基础架构 (NFVI)、VNF 软件及管理和编排 (MANO) 一起来保证服务的可用性。事实上，欧洲电信标准协会 (ETSI) 在其 NFV 弹性规范中声明：“如问题描述中所述，主要目标是确保服务连续性，而不是关注平台可用性。”

## 弹性和可伸缩性

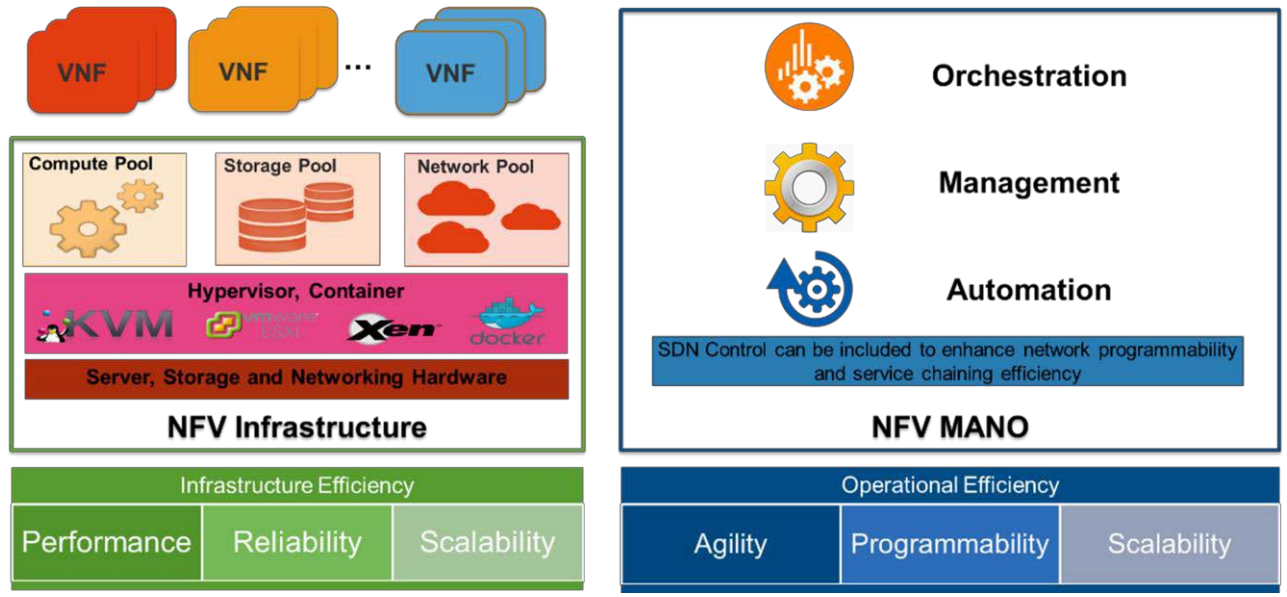
将网络功能转移到云端是 NFV 的最终目标。进入云端后，将可以自动完成服务创建并基于需求进行弹性扩展和缩减。云还可以跨各个位置分发资源，以便将其作为单个资源进行共用和管理，从而帮助实现最大利用率。这对于实现与超大规模云竞争对手相当的基本成本以及相匹配的敏捷性和服务范围尤其重要。

## 性能

不像大多数 IT 应用，电信应用很少是尽力服务。例如，语音和视频应用对延时尤其敏感，经常要求不到 100 毫秒的延迟。并且在电信环境中需要服务数千万的客户，这使性能进一步受到了挑战。

在评估虚拟解决方案时，通信服务提供商期望类似设备具有相似的性能，以便他们可以保持相似的设备覆盖。这意味着对于分组核心演进（EPC）应用，物理服务器可能需要支持 20 Gbit/s 吞吐量，这比典型 IT 应用需要的吞吐量明显更高。

**图 1：NFV 基础架构和 MANO**




*来源：Mellanox*

服务提供商的客户需要高质量的服务（并且一般愿意为其支付额外的费用），并且使用服务级别协议（SLA）来确保其要求得到满足。这些服务器提供商必须确保，当他们转移到虚拟和云化环境时，性能级别需保持不变。

本白皮书的其余部分将重点讨论性能以及可用于支持在虚拟环境中获得平等性能的一些方法。

# 服务器性能的多个维度

在考虑性能时，需要认识到可通过多种方法测量性能，这很重要。

最基本的测量方法是**原始接口吞吐量**，通常以吉比特每秒（Gbit/s）为单位进行测量。这种测量方法描述的是在给定输入/输出（I/O）设备上支持的总流量，例如网络接口卡（NIC）或适配器。如今，10 Gbit/s 网卡很常见，但是要在云中支持 NFV 则需要 40 或 100 Gbit/s。与此同时，与 10 Gbit/s 接口相比，新标准化的 25/50 Gbit/s 接口可多提供 150% 的吞吐量，同时价格也具有竞争力。

同样重要的是**数据包吞吐量**，通常以每秒百万数据包 (Mpps) 为单位进行测量。它描述 I/O 系统处理数据包的速度有多快。许多 VNF 处理大量非常小的数据包，这给系统施加了巨大的压力。在处理小数据包时，保证不丢包要更为复杂和困难的多。

在虚拟环境中，由于 Hyper- visor 层管理的虚拟机和容器可能属于不同的虚拟网络，因此给数据包吞吐量进一步施加了压力。除了虚拟化带来的挑战外，还有服务链的影响。许多 VNF 都由多个虚拟机组成，它们可能在同一台服务器上，也可能不在。不管位置如何，随后的每个虚拟机都会积聚延迟并增加系统开销。如今，大多数服务链都以隧道（例如 GRE/UDP 上的 VXLAN、MPLS）形式实施，从而由于额外的报文头而增加了开销。数据包处理增加了服务链中涉及所有服务器的 I/O 和 CPU 的开销。这些全都会增加延迟，因此，为了实现确定的性能，需要最大程度地减少信号发出和接收开销。

# 克服虚拟化带来的性能冲击

ETSI 在其 NFV 性能规范中承认，“不管 I/O 和内存访问何时与虚拟机实例的整体性能相关，依赖于绕过操作系统的技巧及其 I/O 机制和中断都可能会变得非常重要。” 本部分将介绍一些这方面的技巧。

在这些技巧中使用的技术包括：

- **单根 I/O 虚拟化 (SR-IOV)** 是一项允许多个虚拟机共享单个物理网卡的技术。
- **数据平面开发套件 (DPDK)** 是支持增强数据包处理的一组库和驱动程序。
- **开放 vSwitch (OVS)** 是用于在虚拟机之间路由流量的开源虚拟交换机。它通常驻留在 CPU 的内核空间中。
- **嵌入式交换机（eSwitch）** 是一项驻留在网卡中的新兴技术，它可以加速主机内和主机间数据包的交换。它与虚拟交换机共享转发表，并且预期 eSwitch 最终可能会在某些情况下取代虚拟交换。

## 克服计算虚拟化弊端

如上一部分所述，可能需要多个虚拟机来创建单个服务，并且这些虚拟机驻留在同一或不同服务器的用户空间中。在 Hypervisor 管理的虚拟化环境（相对于裸机）中，虚拟机之间的流量通过 Hypervisor 进入虚拟交换机所在的内核空间。如果进入的是同一个主机，则流量将通过 Hypervisor 返回进入用户空间；如果进入的是不同主机，则会将其放入执行封装、校验和以及 CRC 计算的缓冲区中。在内核中留一份副本，然后将另一份副本发送到 I/O 系统，在这里执行外部数据包封装、校验和以及 CRC 计算。然后，通过网络将流量发送到另一台服务器，在该服务器中反向运行该进程，直到流量到达用户空间中的虚拟机。由于无数的中断和存储转发机制，因此会显著降低数据包的性能。在某些情况下，在 10 Gbit/s 链路上可实现不到 1 Mpps 的性能，理论上会转化为 15 Mpps 的小数据包性能。

### *软件和硬件辅助选项*

已提出多个选项来解决如上所述的性能问题。软件辅助和硬件卸载是其中两个很受欢迎的选择。每一个选择都有其自己的优势和不足。

**软件辅助**可以解决推送模型中出现中断所带来的开销。通过链接 DPDK 库和应用程序编程接口（API），应用程序将连续轮询新数据包，而不是每次有新数据包到达时就中断处理。这种方法的主要优势是它会显著提高处理性能，同时保持硬件的独立性并消除 PCI 带宽开销。但是，除非采用了其他技术，否则仍然会在用户空间发生数据包处理，因此不会降低 CPU 的开销。当 NFV 转向云原生的微服务类型的 VNF 时，可能会进一步加剧此数据包处理开销。

而通过硬件辅助，SR-IOV 可用于启用应用程序直接访问。虚拟机利用直接内存访问（DMA），不再需要将流量复制到缓冲区。可以将这种方法与软件辅助模型中的 DPDK 库结合使用，以获得推送到轮询模式的切换效率。在利用 DPDK 的实现中，经常会将虚拟交换机移出内核，并移入用户空间。在这种方法中，在内核中执行的许多交换功能通常可以在网卡上使用 eSwitch 加速。虽然这种方法增加了 PCIe 负载，但是将交换功能从 CPU 卸载到网卡会加速处理，因为这样可以将 CPU 资源分配给数据包处理以外的其他任务。

在考虑 VNF 时，数据包处理技巧的重要性变得很明显。并不是所有 VNF 都共享相同的要求，例如，有些仅要求数据包终止，而有些也要求数据包处理。后面一种情况将导致有更多东西向流量，向 vSwitche 和网卡施加更多负载，从而使 I/O 成为性能瓶颈。通过使用 SR-IOV 及加速的 vSwitch (AVS)，可以显著提高 I/O 性能，从而使 VNF 软件运行得更接近于本机软件性能。

Affirmed Networks 是一家领先的虚拟化移动解决方案提供商，他们调查了数据包转发体系架构/vSwitch 和网卡吞吐量对其 Mobile Content Cloud (MCC) 软件套件（一个设计为在云环境中运行的虚拟化 EPC 解决方案）整体性能的影响。关于数据包转发，Affirmed 表示 MCC 软件可以按需横向扩展达到服务器的最大 I/O 容量，即等于所有网卡的吞吐量，单网卡和双网卡服务器通常是 10 Gbit/s 或 20 Gbit/s。引入传统 OVS 会使性能降低到服务器 I/O 容量的 10-20%。DPDK 加速的 OVS 可以将实际性能提高到服务器 I/O 容量的 80%，而 SR-IOV 几乎可以达到 100% 的本机线路速度。

另外，Affirmed 发现，随着南北向流量从服务提供商网关路由器进入 EPC（EPC 集群的高效吞吐量），MCC 软件会在 EPC 数据路径中的不同模块之间生成大约相等数量的东西向流量。这意味着为了支持 X 的高效吞吐量，服务器 I/O 容量至少需要达到 2X。例如，为了保持竞争力并支持 20 Gbit/s 的集群容量，MCC 软件将消耗 40 Gbit/s 的服务器 I/O。可通过各种配置支持这一点，随着在以下列表中向下移动，服务器占用面积将逐渐增加，并且计算资源利用率将逐渐减小：

- 一台配有一个 40 Gbit/s 网卡的服务器
- 两台各自配有两个 10 Gbit/s 网卡的服务器
- 四台各自配有一个 10 Gbit/s 网卡的服务器

## 克服叠加 (Overlay) 网络弊端

叠加 (Overlay) 网络技术（例如 VXLAN、NVGRE 和最新的 GENEVE）添加了一个新的报文头和 CRC 来封装流量。这导致给 CPU 资源施加了更大的压力。

传统网卡仅将此报文头看作外部报文头，将数据包从物理源服务器指向物理目标服务器，而不是具有路由信息的内部报文头，将数据包指向运行 VNF 的实际虚拟机或容器。在当前一代的网卡中，许多都只能卸载外部数据包的包校验和/CRC 计算、封装和解封装处理，从而导致处理内部数据包校验和/CRC 计算的 CPU 工作负载增加。

但是，具有叠加 (overlay) 卸载功能的网卡可以处理网卡硬件中的校验和/CRC 计算，从而导致吞吐量明显增加到几乎裸机性能，且用于数据包处理的 CPU 负载显著降低，云基础架构的效率提升。而且，网卡中的 eSwitch 能够同时执行外部和内部数据包封装和解封，并基于内部数据包路由流量。这可以进一步提高吞吐量，并以更低的 CPU 开销获得更具有决定性的延迟。

### 克服 TCP/IP 弊端

联网所用的通信协议 TCP/IP 需要保持状态。TCP/IP 运行于假设有损耗的环境中，并且依赖于丢弃的数据包来作为隐式拥塞通知机制。由此导致的发送方超时、重新传输和无序数据包使卸载 TCP 更加复杂，因此不是卸载，而是通常在 CPU 中进行处理。

RDMA 是一个非专有传输协议，它直接解决了当前计算和联网体系架构的两个重要限制，即在用户/应用程序和内核内存之间进行数据复制产生的开销，以及 TCP/IP 协议引入的延迟。RDMA（在以太网或 InfiniBand 上）设计为提供可靠的有序**消息**序列，而 TCP 设计为提供可靠的有序**字节**序列。它被需要高性能的计算集群广泛采用。

RDMA 基本上属于一个加速的 I/O 交付机制。它引入了“零复制”数据放置概念，允许事务两端专门设计的 RDMA 网卡（也称为 R-NIC）绕过操作系统 (OS) 内核，将数据从源服务器的用户内存直接传输到目标服务器的用户内存。

RDMA 可作为以太网和 IP 的传输协议运行，取代 TCP/IP，从而可以将流量卸载到主机通道适配器 (HCA) 并减少 CPU 开销。这一过程的工作方式与如上所述 DPDK 支持的硬件卸载相似，除了它使用的是 Accelio（开源消息传递和远程过程调用库）。

## 结论

离开专用硬件平台转向标准化 IT 服务器是伴随 NFV 所带来的最显著的变化之一。不仅这些服务器未设计为提供通信服务提供商所需的可靠性、弹性和可伸缩性，而且诸如计算虚拟化、叠加 (overlay) 和 TCP/IP 等技术也增加了性能损失。

为了克服这些性能问题，可以使用软件和硬件加速技术。利用 DPDK 可以降低 CPU 开销，而 SR-IOV 则为虚拟机提供了共享单个物理网卡的方法。此外，新网卡有助于解决叠加 (overlay) 网络问题，同时 RDMA 的出现成为可替代 TCP/IP 的传输方案。

## 关于 Mellanox

Mellanox 提供广泛的以太网和 InfiniBand 适配器。它最先推出了 40 Gbit/s 网卡，并且最近推出了支持 100 Gbit/s 和新的 25/50 Gbit/s 规格的网卡。Mellanox 的适配器支持 SR-IOV 和 OpenFlow 实现的 eSwitch 功能，可提供硬件加速，克服伴随虚拟化及 VXLAN、NVGRE、GENEVE 和 MPLS 叠加 (overlay) 网络而来的性能问题。它们也在融合以太网上支持 RDMA，以克服 TCP/IP 固有的性能损失。